DOCUMENT RESUME

ED 093 985 TM 003 841

AUTHOR Follman, John; And Others
TITLE Kinds of Keys of Student Ratings of Faculty Teaching Effectiveness.
PUB DATE [Apr 74]
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS College Students; *College Teachers; *Effective Teaching; Measurement Techniques; *Rating Scales; Reliability; *Response Mode; *Teacher Rating; Testing Problems

ABSTRACT

Three substudies of effects of different formats on student ratings of faculty teaching effectiveness were conducted. One substudy investigated Kinds of Keys, Agreement, Evaluation, and Needs Improvement. The second, NO TUP, (New Observation of Teaching of University Professor Rating Scale), investigated numbers of positive rating categories. The third, Wording, investigated the same items worded positively, negatively, and neutrally, respectively. Practically important differences in level of ratings obtained in Kinds of Keys, and practically and statistically significant differences obtained in NO TUP and Wording. Additional research is necessary to determine if apparent differences in teaching effectiveness are actually differences in teaching effectiveness or differences in the methods of measurement. (Author)

ED 093985

BEST COPY AVAILABLE

Kinds of Keys of Student Ratings of Faculty Teaching Effectiveness

John Follman Manny Lucoff Leslie Small Fred Power

University of South Florida

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

INTRODUCTION

There currently is considerable interest in college faculty, administrators, and students in student evaluation of the effectiveness of college courses and college professors. The most commonly used means of obtaining student evaluation of instruction is student ratings. While there has been considerable interest in student evaluation there has not been a corresponding amount of research particularly on the technical aspects of student rating scales. Some of the technical aspects on which little research has been reported are the keys and the formats.

There has been little applied research reported on the effects of kinds of keys and forms of formats used in student rating scales of faculty teaching and course effectiveness. However there is relevant basic research on kinds of keys and formats in rating scales in job performance ratings. Barrett, Taylor, Parker, and Martens (1958) investigated four formats: trait names only: verbal definitions of traits: trait names and behavioral descriptions but no definitions; and trait definitions and behavioral descriptions but no trait names. There were significant differences for formats and for all interactions involving formats. Higher ratings were associated with trait names and behavioral descriptions but no definitions. In a similar study Madden and Bourdon

TM 003 841

(1964) examined various format forms including horizontal, vertical, bars, no bars, number, and labels arrangements. Again there were significant differences for formats and also all interactions. It is apparent from these basic rating research studies that the methods of measurement, as well as the variables measured, influence the level of ratings awarded.

The purpose of this paper is to report the results of three substudies of kinds of keys for college student ratings of college professors' teaching effectiveness. In the first substudy, Kinds of Keys, the three main kinds of keys were investigated: Agreement; Evaluation; and Needs Improvement. The reason for this specific substudy was to determine if the different rating contexts per se influenced the level of ratings awarded.

In the second substudy, NO TUP, four sets of evaluative keys ranging from two negative, one neutral, two positive, to all five positive categories, were investigated. NO TUP is an acronym for the New Observation of Teaching of University Professors rating scale. A response set that characterizes many rating situations is the leniency (generosity) effect. The leniency effect is the tendency of raters to consistently assign ratings that are too high. In order to ameliorate this problem Guilford (1954) recommended an unbalanced set of categories with three positive, one neutral, and one negative rather than a conventional set of categories with two positive, one neutral, and two negative. In another milieu involving ratings, essay grading of English compositions, Follman and colleagues conducted three relevant studies (Follman and Reilly, In press; Follman, 1972; Follman, Silverman, and Reilly, 1972). In these three analyses the following kinds of categories were investigated: Numbers (5, 4, 3, 2, 1); Negative categories (one positive, one neutral,

three negative): Conventional (two positive, one neutral, two negative); and Guilford categories (three positive, one neutral, one negative). Across all three studies all sets of categories were reliable. Across all three studies it was concluded that kinds of categories influence level of ratings, that negative categories produce the highest ratings, and that positive (Guilford) categories produce the lowest ratings. Thus the reason for the second specific substudy was to determine if different combinations of positive categories would reduce the leniency error in student ratings of instructor teaching effectiveness as they did in English composition scoring.

In the third substudy, Item Wording Direction, the agreement keys used in the Kinds of Keys substudy were used for the same set of items each set respectively worded positively, negatively, or neutrally. Two basic rating research studies were identified in which item phrasing was varied positively and negatively. Whipple (1957) compared positive and negative phrasing in an item writing study. Little differences were found between the two forms of phrasing but there was a tendency for "true" to be given to positively worded items. Ishikawa (1966) made a number of empirical comparisons including one between affirmative statement and question statement formats and found few differences. Thus the reason for the third specific substudy was to determine in the context of student rating scales if the item wording tone of different rating sets, positive, negative, or neutral, would influence the level of ratings awarded. The three studies are depicted in Table 1.

The objective across all three substudies was to determine if the keys per se affected the level of student ratings of faculty teaching effectiveness. Reliability was also considered, but it was not anticipated

Table 1

Depiction of the Three Substudies

Substudy #1 Kinds of Keys				
Evaluation				
Unsatis-factory	Below Average	Average	Above Average	Excellent
1	2	3	4	5
Agreement				
Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
1	2	3	4	5
Needs Improvement				
Needs Improve-ment	Needs Consid-erable Improve-ment	Is Adequete	Needs Little Improve-ment	Needs No Improve-ment
1	2	3	4	5

Substudy #2 NO TUP Conventional (Evaluation)				
Unsatis-factory	Below Average	Average	Above Average	Superior
1	2	3	4	5
Garden Variety				
Below Average	Average	Above Average	Superior	Excellent
1	2	3	4	5
Nirvana				
Average	Above Average	Superior	Excellent	Superb
1	2	3	4	5
NO TUP				
Above Average	Superior	Excellent	Superb	Perfect
1	2	3	4	5

Substudy #3 Item Wording Direction*
Positive Wording
Instructor knows the material well.
Negative
Instructor knows the material poorly.
Neutral
Instructor knows the material moderately.

*The five category agree-disagree key was used for all three formats.

to be critical because of the substantial size of each treatment group within each sample within each substudy.

PROCEDURE

All three substudies were conducted in December 1972 at the University of South Florida. The Ss for substudy #1 (Kinds of Keys) were students in an undergraduate finance course. The Ss for Substudy #2 (NO TITLE) were undergraduates in another section of the finance course conducted by another instructor. The Ss for Substudy #3 (Item Wording Direction) were students in an undergraduate broadcasting course. The three instructors were chosen because they were considered to be characteristic college teachers. Operationally, this means that they usually receive student rating near four on a five point scale.

The rating scale for Substudy #2 consisted of 17 conventional college teaching effectiveness rating items developed at the University of South Florida. These items were also used for Substudy #1 and Substudy #3. In addition, 23 items, developed by the University of South Florida College of Education, were also used.

The students in the class composing each respective substudy were randomly assigned to its respective treatment conditions.

The ratings were quantified 5, 4, 3, 2, or 1 for the statistical analyses as indicated in Table 1.

ANOVA adjusted group reliability estimates were determined for each treatment group for each substudy.

Means, standard deviations, and ANOVA's were computed for total of items, and individual items, for each format to determine the effects of each treatment format within each substudy on level of ratings.

RESULTS

Table 2 indicates the group reliability estimates, means adjusted to the five point scale, and total score means and standard deviations for each treatment group within each substudy, across all three substudies.

Initially, it is apparent from Table 2 that all treatment groups across all three substudies rated reliably. Even the Item Wording Direction Neutral group's estimate, .79, the lowest, is adequate. It is likely that this group's estimate would have been higher had its size been bigger. Consequently considerable confidence can be placed in the integrity of each group's ratings as a dependent variable in the three subsequent treatment effects analyses.

The treatment effect analyses will be treated sequentially. For Substudy #1 Table 2 indicates a lower mean for the Evaluation format vis a vis the Agreement and Needs Improvement formats. An ANOVA indicated a non-significant F of 2.20 for 2 and 106 degrees of freedom. One-way ANOVA's for each individual item indicated significant differences for keys for only one item. Despite the general non-significant nature of these differences it is suggested that an absolute rating difference of .23 or .25 (3.71 vis a vis 3.94 and 3.96) on a five point scale would not be perceived indifferently by faculty of an institution where such ratings were used administratively. This is particularly true because the Evaluation format is used much more frequently than the Needs Improvement format. Therefore it is suggested that additional research be carried out on this issue using a large number and variety of instructors and students.

Examination of the results for Substudy #2, NO TUP, indicates some fascinating findings. Specifically, the adjusted means were 4.13, 3.69,

Table 2

N's, Reliability Estimates, Adjusted Means, and Total Means and Standard Deviations for Each Treatment Group Within Each Substudy, Across All Three Substudies

Substudy #1 Kinds of Keys						Substudy #2 NO TUP						Substudy #3 Item Wording Direction					
	N	Rel.	Adj. $\bar{X}$	$\bar{X}$	SD		N	Rel.	Adj. $\bar{X}$	$\bar{X}$	SD		N	Rel.	Adj. $\bar{X}$	$\bar{X}$	SD
Eval.	33	.96	3.71	148.33	22.39	Convent. (Eval)	25	.90	4.13	70.32	8.76	Pos.	20	.98	3.36	134.55	31.45
Agree.	40	.96	3.94	157.50	20.39	Card. Var.	27	.94	3.69	62.89	13.38	Neu.	22	.79	3.03	121.23	13.12
Improve.	36	.96	3.96	158.39	23.36	Nirv.	27	.93	3.63	62.55	12.92	Neg.	19	.96	2.15	85.90	27.86
						NO TUP	27	.95	3.51	59.59	15.79						

3.68, and 3.51, respectively, for Conventional, Garden Variety, Nirvana, and NO TUP, respectively. The ANOVA indicated a significant ($p < .05$) F of 3.15 for these keys. ANOVA's for individual items employing a conservative level of significance indicated three (of 17) items significant ($p < .05$). Since the means range significantly from 3.51 to 4.13 depending upon the kinds of categories used, and since this range could be extended even more by using negative categories it is clear that the kinds of categories used influence the level of ratings awarded. The corollary conundrum is the issue of which categories to use. If the assumption is made that college professors in general are better teachers than teachers in general (the focal argument would be that they know more) then some set of categories employing more positive than negative categories would probably be appropriate. If, on the other hand, the assumption is made that college professors are not better than teachers in general or else that student raters should compare the particular professor with other professors only and not with teachers in general, then a balanced set of categories might be appropriate. In any case it is clear that kinds of categories to be used in rating instructors is an issue that should be considered seriously.

Examination of the results for Substudy #3, Item Wording Direction, indicates means of 3.36, 3.03, and 2.15, respectively, for Positive, Neutral, and Negative categories, respectively. The ANOVA indicated a highly significant ($p < .001$) F of 19.6 for 2 and 38 degrees of freedom for these rating set tones. ANOVA's for individual items again using a conservative significance procedure indicated 21 (of 40) items significant ($p < .05$). These findings are viewed as additional evidence of the effects of format factors in addition to the actual competence of the instructor being considered. It is not considered that these findings are otherwise important,

for two reasons. One reason is that in order to make this substudy similar structurally to the other two substudies some conceptual interpretative uncertainties were built into the combination of the individual items and the agreement format. Secondly, hopefully no one will employ a negative format.

OVERVIEW

Overview of these three substudies indicates the following conclusions.

Initially, it is compellingly clear that kinds of categories influence massively level of ratings awarded.

Secondly, it is also compellingly clear that this source of spurious variance will have to be taken into account in any administrative application of student ratings of faculty teaching effectiveness. The paramountcy of this issue is evident when it is considered that most faculty fall within 1.5 ratings on a conventional five category scale and that NO TUP alone manipulated .62 of a rating unit, almost half of the actual functional range from which to differentiate faculty, assuming such an administrative objective.

Third, additional research is recommended on the kinds of formats employed, i.e., Evaluation, Agreement, Needs Improvement. It appears that while there may be some limited differences in level of ratings awarded it might be prudent to please faculty by using either the Agreement or Needs Improvement formats particularly if the rating levels are similar.

Fourth, both empirical evidence and rational research need to be reported on the question of which kind of categories should be used. This enigma has philosophical implications on the number system to be used in quantifying the data for the statistical analyses.

Fifth, while rating reliability is not a matter of concern validity is a vital concern. Implicit in the use of student ratings is the assumption that students are an appropriate, valid public. This is probably more of a normative question than an empirical one considering the characteristic correlations of .30 - .40 between student ratings of teacher effectiveness and the students' achievement (Follman, 1972).

Finally, the pro forma caveat is noted that the results reported herein are veridical to the extent that the three instructors used represent college instructors in general. It is considered that the total item ratings of 3.71, 4.13, and 3.36, respectively, for Instructors #1, #2, and #3, respectivley, on the most conventional keys are certainly at worst ball park figures. The high reliability estimates provide additional support for this interpretation. In any case it is recommended that research be conducted on the questions raised herein as they are integral to any administrative application of student ratings.

REFERENCES

Barrett, R.S., Taylor, E.K., Parker, J.W., and Martens, L. Rating scale content: Scale information and supervisory ratings. Personnel Psychology, 1958, 11, 333-346.

Follman, J. Essay grading formats. Unpublished Study, University of South Florida, 1971.

Follman, J., and Reilly, R. Generosity in essay grading. Florida Journal of Educational Research, 1973, 15, 79-82.

Follman, J., Silverman, S., and Reilly, R. Generosity in grading formats revisited. Florida Educational Research Association, Tampa, January, 1973.

Follman, J. Student ratings and student achievement. Unpublished article, University of South Florida, 1972.

Guilford, J.P. Psychometric Methods, New York: McGraw-Hill Book Company, 1954.

Ishikawa, A. A study of the effects of different forms of questionnaire items on factorial validity. Psychologia, 1966, 9, 76-84.

Madden, J.M., and Bourdon, R.D. Effects of variations in rating scale format on judgment. Journal of Applied Psychology, 1964, 48, 147-151.

Whipple, J.W. A study of the extent to which positive or negative phrasing affects answers in a true-false test. Journal of Educational Research, 1957, 51, 59-63.